Abstract No: SO-05

Capturing sentence-level positional data into N-gram profiles for document classification

L. M. S. Gunasekara[1*] and H. K. S. Premadasa[2]

[1]Department of Statistics & Computer Science, University of Kelaniya, Sri Lanka, [2]Sabaragamuwa University of Sri Lanka
gunaseka_ps17050@stu.kln.ac.lk*

Document classification is a crucial aspect in natural language processing with a wide range of applications in various domains such as email spam filtering, hate speech detection, political bias assessment, etc. While modern *transformer-based* classification approaches have shown promising results in this area, they rely on expensive parallel processing hardware, leaving them out of reach for simpler applications. Therefore, it is still safe to assume that there is room for improvement in terms of developing approaches with lower computational complexity. *N-grams* are a simple and efficient way of representing text data as features based on the distribution of contiguous tokens within the text. This approach is widely used in text analysis and research due to its language independence and minimal pre-processing requirements. However, most of these models do not possess sentence-level positional information in their *n-gram* profiles. Hence, in this study, we propose a revised algorithm for generating *n-gram* profiles related to document categories in a classification task. We combine this new algorithm with the *Euclidean* distance metric to assign class labels for raw documents. This algorithm was evaluated on two main tasks: language classification and subject classification (in English). Our results show that this approach achieves accuracy levels comparable to state-of-the-art models. For the language classification task, we were able to showcase an accuracy of 91% on the WiLI Benchmark Dataset consisting of 235 languages in total with an average prediction time of 1.88×10^{-2} seconds. Furthermore, we investigated several configurations in the dimensions of *n-gram range* and *n-gram cutoff length* for the subject classification task. The best performing configuration of a fixed n-gram length of 5 and a cutoff length of 5000 assumes an accuracy of 50% with an average inference time of 3.29×10^{-2} seconds on the 20 Newsgroups Dataset spanning a whole of 20 newsgroups categories. Overall, our findings suggest that this approach of including sentence-level positional data in *n-gram* profiles can facilitate an algorithm of minimal complexity, and this algorithm, combined with a suitable *n-gram range* and *cutoff level,* can perform well for document classification, particularly when dealing with noisy data with similar categorical labels.

Keywords: Document Classification, N-Grams, Natural Language Processing, Language Classification, Subject Classification